Dornburg et al. BMC Evolutionary Biology (2016) 16:262
DOI 10.1186/s12862-016-0837-3

BMC Evolutionary Biology

SOFTWARE

Open Access

PhyInformR: phylogenetic experimental design and phylogenomic data exploration in R

Alex Dornburg[1*], J. Nick Fisk[2], Jules Tamagnan[3] and Jeffrey P. Townsend[2,4,5]

Abstract

Background: Analyses of phylogenetic informativeness represent an important step in screening potential or existing datasets for their proclivity toward convergent or parallel evolution of molecular sites. However, while new theory has been developed from which to predict the utility of sequence data, adoption of these advances have been stymied by a lack of software enabling application of advances in theory, especially for large next-generation sequence data sets. Moreover, there are no theoretical barriers to application of the phylogenetic informativeness or the calculation of quartet internode resolution probabilities in a Bayesian setting that more robustly accounts for uncertainty, yet there is no software with which a computationally intensive Bayesian approach to experimental design could be implemented.

Results: We introduce PhyInformR, an open source software package that performs rapid calculation of phylogenetic information content using the latest advances in phylogenetic informativeness based theory. These advances include modifications that incorporate uneven branch lengths and any model of nucleotide substitution to provide assessments of the phylogenetic utility of any given dataset or dataset partition. PhyInformR provides new tools for data visualization and routines optimized for rapid statistical calculations, including approaches making use of Bayesian posterior distributions and parallel processing. By implementing the computation on user hardware, PhyInformR increases the potential power users can apply toward screening datasets for phylogenetic/genomic information content by orders of magnitude.

Conclusions: PhyInformR provides a means to implement diverse substitution models and specify uneven branch lengths for phylogenetic informativeness or calculations providing quartet based probabilities of resolution, produce novel visualizations, and facilitate analyses of next-generation sequence datasets while incorporating phylogenetic uncertainty through the use parallel processing. As an open source program, PhyInformR is fully customizable and expandable, thereby allowing for advanced methodologies to be readily integrated into local bioinformatics pipelines.
Software is available through CRAN and a package containing the software, a detailed manual, and additional sample data is also provided freely through github: https://github.com/carolinafishes/PhyInformR.

* Correspondence: alex.dornburg@naturalsciences.org
[1]North Carolina Museum of Natural Sciences, Raleigh, North Carolina 27601, USA
Full list of author information is available at the end of the article

Background

The 21st century has witnessed rapid progress in the phylogenetic resolution of the evolutionary relationships in the Tree of Life, associated with a trend toward analysis of multi-locus and even genome-scale datasets [1–4]. However, despite this wealth of data, achieving resolution of some key nodes in the tree of life remains challenging. Some nodes persistently elude any resolution; others are characterized by conflicting results, each based on different sets of data, and each well-supported by current metrics of support [5, 6]. A critical step towards resolving and stabilizing our understanding of the Tree of Life is the development of tools that identify potential sources of non-phylogenetic signal—parallelism or convergence in character state that do not reflect shared evolutionary history—on phylogenomic datasets [5, 7].

Analyses of phylogenetic informativeness (PI; [8]) and quantification of quartet internode resolution probabilities (QIRP) [9, 10] provide a framework with which to predict the phylogenetic utility of large multi-locus and phylogenomic datasets. These analyses provide insights into the predictive utility of sequences across entire topologies as well as for individual nodes. For example, the shape of PI profiles can be used to determine either the potential utility of data for inference or the severity to which phylogenetic information content in a dataset has decayed over the temporal history of a clade [8, 11]. Although a useful heuristic, PI profiles are based solely on rates of character change and therefore provide no direct prediction of how homoplasious site patterns will influence phylogenetic resolution of specific nodes [11, 12]. However, QIRP calculations make use of theory that posits a predictive relationship between the rate of evolution, specified internode distances, and tree depth with the probability of resolving a given node thereby allowing for a more detailed prediction of how homoplasy will impact inference [9].

Using a rate of sequence evolution for an empirical dataset in conjunction with an s-state Poisson model ($s \geq 2$), Townsend et al. [9] derived a model that calculates the probability of resolution for a given phylogenetic quartet. These calculations quantify not only the probability of correct resolution (QIRP), but also quartet internode homoplasy probabilities (QIHP) that represent the probability of having greater strength of support at a given internode for an incorrect rather than correct quartet topology as well as quartet internode polytomy probabilities (QIPP) that represent the probability of no resolution [9]. These calculations can empower sequencing pipelines to generate data that will resolve specific problems. Specifying a range of possible tree depths and internodes with estimated rates of markers, investigators can rank the effectiveness of candidate loci, saving both time and sequencing costs. Although the development of these methods targeted phylogenetic experimental design prior to sequencing [8], they have been successfully applied as data filtration metrics [13] and used to assess the validity of inferences based on existing markers [3, 14]. This additional utility makes these calculations potentially very useful in bioinformatic pipelines aimed at selecting loci from existing genomic datasets for analyses. However, software enabling the application of these tools is extremely limited.

The only tools currently available for analyses of phylogenetic informativeness are the locally implementable program TAPIR [15], which only generates PI profiles, and the PhyDesign web application [16], which has a modest server-based computational throughput and a limited scope of functionality. While these applications are useful, their limitations are stymieing adoption of phylogenetic informativeness and quartet probability based methods. There are powerful theoretical advances not reflected in these applications. For example, quantification of QIRP or QIHP in existing software assumes an underlying Jukes-Cantor nucleotide substitution matrix, while theory has been derived that facilitates a more accurate calculation that can be conducted with any substitution model [17]. Furthermore, existing applications force an assumption of equal divergence times for all four taxa in a phylogenetic quartet. This assumption is commonly not met in empirical datasets [18], and theory has been derived that facilitates calculations for any set of branch lengths [19]. A third major limitation of current implementations is that there is no way to integrate over branch length uncertainty using Bayesian posterior distributions. Finally, the ability to visualize results with previous tools is modest and does not meet the needs of investigators with phylogenomic or next-generation sequence data.

Publicly available software that addresses these three criteria is a critically needed resource for assessing the utility of potential or actual data sets for phylogenetic inference. However, open source software facilitating both advanced calculations and visualizations of information content has been lacking. Here we present the software package PhyInformR, an open source program—designed to keep pace with advances in theory—that will allow users to quantify and display the phylogenetic information of computationally demanding datasets.

Implementation

PhyInformR is an open source software package in R [20] that utilizes the phylogenetic packages APE [21], PhyTools [22], and Geiger [23]. As the expanded theory enabling any substitution model and uneven quartet trees [17, 19] require solving symbolic equations. To facilitate faster throughput, parallel processing is supported using the foreach software package [24]. PhyInformR can be downloaded along with an in depth user guide and example phylogenies [25–27] from

github: https://github.com/carolinafishes/PhyInformR. As an input, PhyInformR uses site rates estimated from software such as Hyphy [28] and user tree topologies to rapidly enable quantification of PI across datasets or user defined dataset partitions.

Rather than being a black box, the flexibility of the R language allows users to rapidly define any subsets of their data to quantify all metrics available in the phyDesign web interface [16] such as PI profiles [8] or QIRP, QIPP, or QIHP values [9]. PhyInformR expands the power of the phyDesign web interface [16] by incorporating recently published theory that calculates metrics on user specified quartets with uneven branch lengths [19] using any specified symmetrical substitution model and any empirical distribution of base frequencies [17]. The software also accommodates phylogenetic uncertainty, enabling these calculations to be performed across Bayesian posterior distributions of user supplied tree topologies and branch lengths. To overcome challenges in the visualization of information content in phylogenomic datasets, PhyInformR additionally offers a flexible suite of options (Additional file 1) for users to view quantified values of QIRP, QIPP, or QIHP that utilize customizable graphics packages such as ggplot2 [29].

Results and discussion

Expanding computational throughput

We used the avian tree topology from Prum et al. [3] and corresponding site rates estimated using Hyphy [28] to quantify and plot PI profiles [16] (Fig. 1a). PI profiles of the concatenated data demonstrate a slow decline towards the root of the tree, indicative of increased homoplasy [11, 13]. By defining two partitions based on two categories of site-rates, visualizations of PI profiles demonstrate the limited utility of faster evolving sites for deeper nodes (Fig. 1b). Explorations of this type, such as additional evaluations of individual loci or automatically generated partitions of site rate vectors by codon position (provided the generating alignment was in frame) allow for both dynamic adjustments of plots and assessing the impact of site removal on PI profiles. This type of interactive data exploration was not previously possible and is of particular utility for estimation of divergence times. Finding dataset partitions that minimize decay in PI profile has been suggested to aid in mitigating the impacts of branch length estimation errors in divergence dating analyses [13]. Given the computational expense of estimating divergence times using phylogenomic scale data, identifying loci with the lowest declines of PI offers a potential data selection criteria for divergence time analysis pipelines [3].

Although PI profiles provide a useful visualization, it is important to note that PI profiles perform no explicit quantification of how homoplasy will impact tree topology [11]. Prediction of the impact of homoplasy can be conducted by calculation of QIRP and QIHP [8, 9, 17]. Due to the computational demands of the Monte Carlo approach to solving for the probabilities presented in Townsend et al. [9], the phyDesign web interface enables download links only for user-defined nodes of interest, and only computes results using a closed-form analytical approximation. This approach does not allow for interactive exploration of data partitions, and also does not yield the more nuanced graphical output of the higher moments of the resolution probability distribution [16]. In contrast, PhyInformR performs both the rapid analytical approximation and the intuitive Monte Carlo visualization (encoded to rapidly execute in parallel [24], if desired). For example, using the two partitions of Prum et al. [3] data examined above, we can compare the predicted probability distribution of the slower versus faster site rates for the most recent common ancestor of *Opistocomus hoatzin* and other birds. In this case, the slower site rate partition provides a

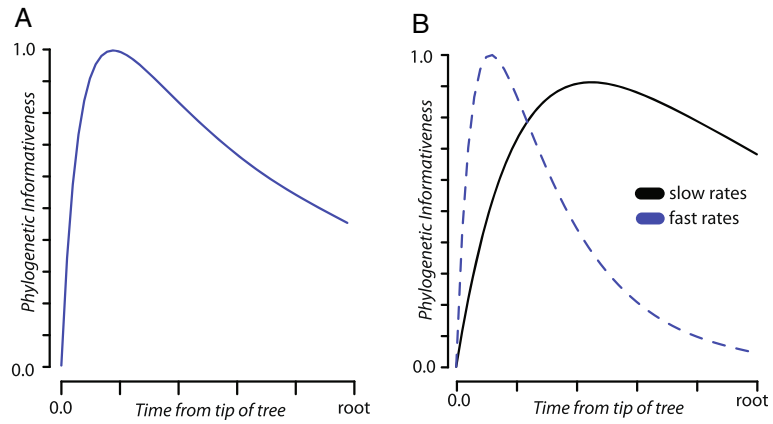

Fig. 1 Phylogenetic informativeness profiles generated in PhyInformR based on a selection of 60 loci from [3] for (**a**) all loci; (**b**) a user defined partition of "fast" versus "slow" rates of nucleotide substitution, with the upper 10% of the rate distribution selected as "fast"

higher probability of correctly resolving this node (Fig. 2a). Additionally, the fast site partition demonstrates a wider spread of potential support for the correct or incorrect topology, conveying both the potential for high support for a correct result and the elevated risk of spurious results when using this partition in phylogenetic analyses (Fig. 2b).

Advanced calculations, Bayesian integration, and result visualization

Early implementations [9] assume a Jukes-Cantor model of substitution [30] as well as equal branch lengths in the quartet tree. Estimated model fits and tree topologies from empirical datasets suggest that these assumptions are commonly violated [18, 31–33]. PhyInformR encodes new theory allowing users to specify any symmetrical substitution model, and to specify uneven branch lengths in the quartet. Furthermore, PhyinformR can incorporate uncertainty in tree topology and branch lengths into these calculations. This integration over uncertainty of the underlying phylogenetic tree provides a major advantage in comparison to available software that only permits users to evaluate single tree topologies with fixed branch lengths [16]. Use of a single tree can be problematic since quantifications of QIRP, QIHP, and other PI based metrics are sensitive to topology and branch lengths [3]. PhyInformR enables users to iterate calculations over any set of hypothetical trees or over posterior distributions from Bayesian phylogenetic analyses. These parallelizable calculations enhance the robustness of QIRP, QIPP, and QIHP.

The ability to store these calculations into objects creates a foundation for numerous advances in visualization. For example, visualizing QIRP values across an entire tree for multiple dataset partitions simultaneously [34] allows for the assessment of whether a concatenated dataset with an overall low QIRP may contain several partitions with high QIRP that could be assembled to resolve a specific node. In the case of the Prum et al. avian dataset [3], we can see that one partition of the dataset has higher predicted utility for resolving more of the deep internodes of the avian Tree of Life, whereas another has much lower probabilities (Fig. 3a). To account for uncertainty in internode lengths, heatmaps can be generated to provide a graphical display of information content across a range of internode lengths, portraying trends in utility. For example, heatmaps of individual loci in the avian dataset (Fig. 3b) reflect a trend of decreasing signal at small internodes. These sorts of trends shed light on sources of gene-tree conflict, illuminating whether lack of power (high QIPP) or any combination of convergence and parallelism (high QIHP) are driving discordance. Further, calculation of QIRP, QIPP, and QIHP across posterior distributions of trees enables visualizations that convey how the predicted power of data to resolve a node changes with the degree of uncertainty present for a focal phylogenetic quartet topology and internode lengths. For example, an analysis of the RAG1 gene from a study of Bichirs [27] depicts nearly equal kernel probability densities between QIRP and QIPP for a node that was—indeed—not strongly supported in the original study (Fig. 4a). A frequency plot of internode lengths for each measure demonstrates that QIRP decreases at small internodes, while both QIPP and QIHP increase (Fig. 4b). Not only does the probability of homoplasy increase with shorter internode lengths, but small differences in the branch length of increasingly small internodes exerts increasingly larger effects on the predicted utility of a dataset (Fig. 4c & d). In this example, these results nevertheless lead to the expectation that the attempt to confidently resolve this particular node is underpowered.

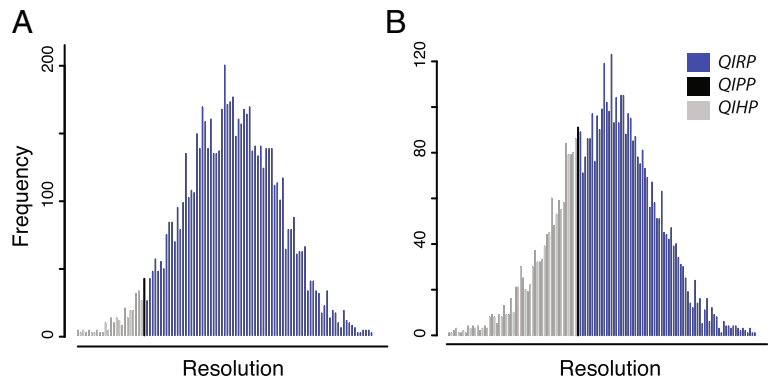

Fig. 2 Graphical output of the Monte Carlo based analysis of QIRP (*blue*), QIPP (*black*), and QIHP (*grey*) shown from right to left respectively. Quantification was based on a selection of 60 loci from [3] for the most recent common ancestor of *Opistocomus hoatzin* and other birds. **a** Results from the slower site rate partition depicting a higher probability of correctly resolving this node. **b** Results from a fast site partition demonstrating a wider spread of potential support for the correct or incorrect topology, depicting the elevated risk of spurious results from this partition

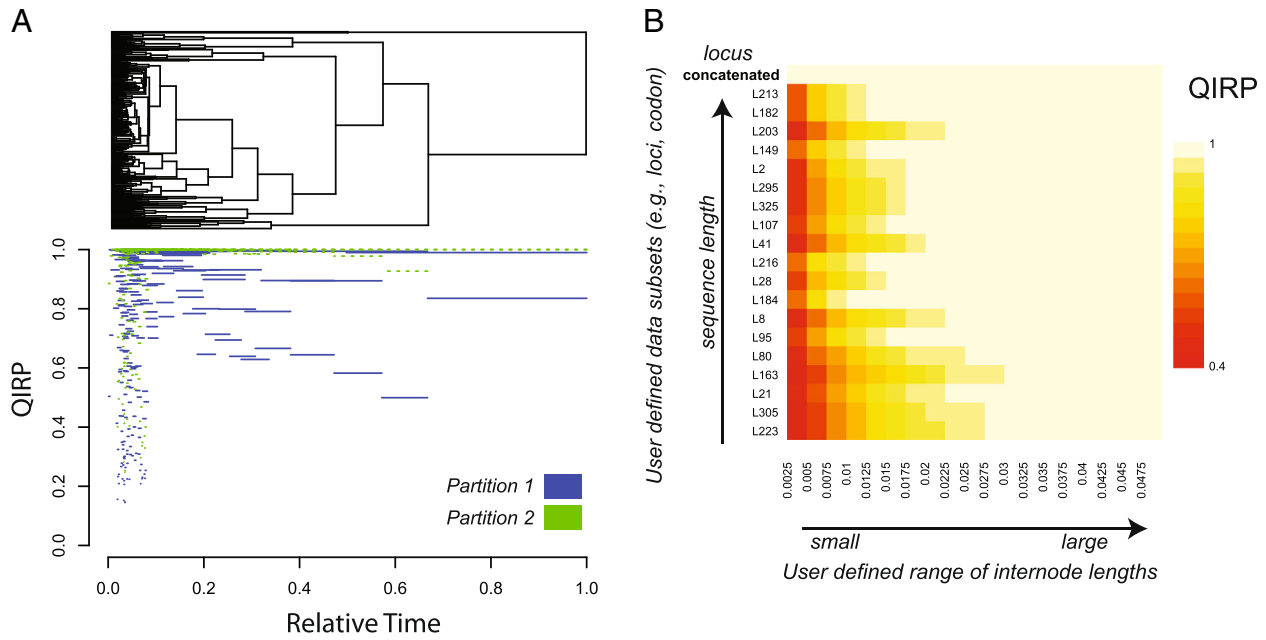

Fig. 3 a Visualizations of QIRP as in [20], comparing QIRP calculations of two dataset partitions from [3] across the tree. Partition one represents "slower" evolving sites based on a cutoff of 0.003 subs/site per million years while partition two represents "faster" evolving sites that lie above the user defined cut off value. **b** QIRP heat map from [3] with probabilities of QIRP quantified across a range internode lengths for a shallow node. Each row corresponds with a user defined delimitation of site-rates (loci and a concatenated dataset in this example, though heatmaps can be modified to any user defined partitions). Shadings correspond with QIRP values. Arrows in the x and y axes respectively indicate internode or sequence length from small to large

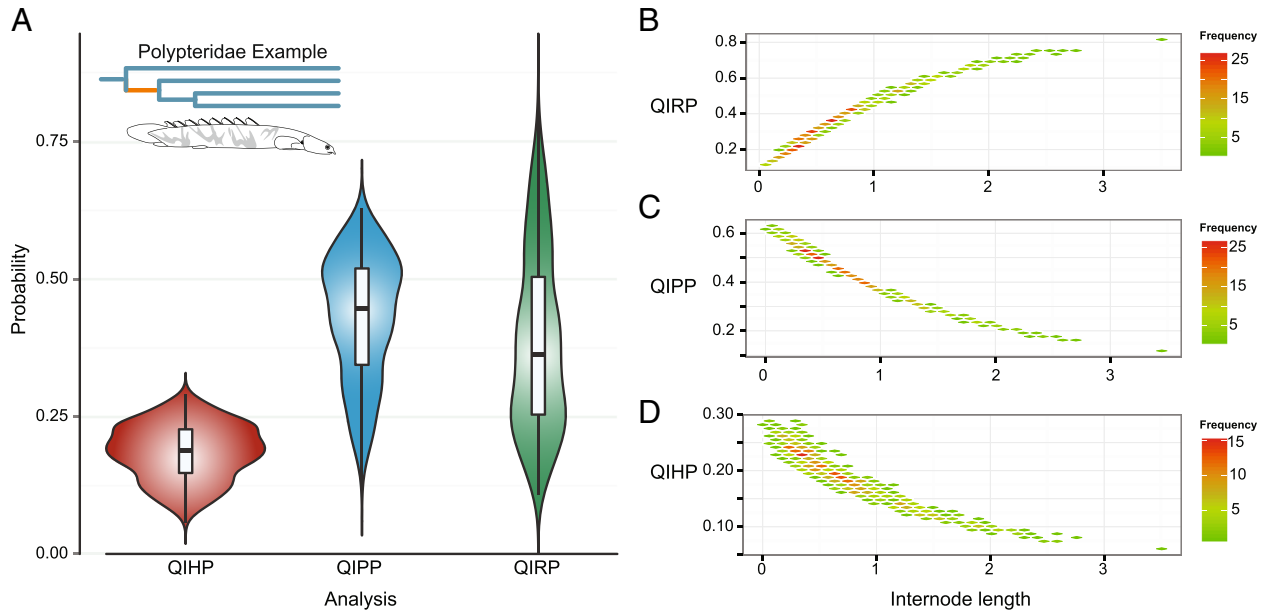

Fig. 4 Two visualizations of Quartet internode resolution (QIRP), polytomy (QIPP), and homoplasy (QIHP) probability calculations in PhyInformR using a focal node of bichirs (most recent common ancestor of *Polypterus congicus* and *P. ansorgii)* with Bayesian time trees and RAG1 sequence data from Near et al. [18]. **a** Violin plots displaying the kernel probability density of each of the three calculations across the posterior distribution of trees for the focal node with box plots of the quartiles for each distribution overlaid within. **b**–**d** Frequency plots for comparing QIRP (**b**), QIPP (**c**), and QIHP (**d**) against the length of internode. All points are sampled from the posterior. Color indicates the frequency of the joint QI measure and internode length within the posterior. In the Near et al. [18] tree from which this internode came, this internode was unresolved; this plot demonstrates that in most samples from the posterior, internode length was short, conferring little power to resolve branching order

Conclusions

PhyInformR provides a new toolset in the phylogenetic toolbox that characterize phylogenetic information in next-generation sequence datasets, enabling both new approaches to experimental design and dataset scrutiny. For targeted phylogenetic studies, PhyInformR will allow groups of loci to be screened for their phylogenetic utility prior to sequencing, potentially cutting costs and time. Likewise, screening loci during probe set development in next-generation sequencing bioinformatic pipelines can cut sequencing costs and the necessity for data filtration during downstream analyses. Furthermore, PhyInformR can complement investigations of topological or branch length incongruence, and can provide insight into sources of error, in some cases facilitating conclusive resolution of nodes that would otherwise remain contentious. The flexibility of graphical output in R makes PhyInformR an expansible tool set for dataset exploration allowing continued development of approaches to visualizing trends in genome-scale datasets. Such capabilities are critical not only to better our understanding of sources of topological incongruence, but also to the goal of continuing to increase our ability to resolve a robust and accurate Genomic Tree of Life.

Availability and requirements

PhyInformR is implemented in R with the package available on CRAN and at: https://github.com/carolinafishes/PhyInformR

Additional file

Additional file 1: PhyinformR user guide and tutorial. (PDF 7055 kb)

Abbreviations

PI: Phylogenetic informativeness; QIHP: Quartet internode homoplasy probability; QIPP: Quartet internode polytomy probability; QIRP: Quartet internode resolution probability

Acknowledgements

This research was supported by an award from the Notsew Orm Sands Foundation to JPT. We thank R. Etter and A. Lamb for help with the graphic design. We also thank two anonymous reviewers for their help and suggestions for improvement to both the software and the manuscript and T. Su for help with Mathematica.

Authors' contributions

AD and JPT conceived of and designed the project. AD, JT, and JNF coded software. AD drafted the initial manuscript. AD and JPT revised the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approvals and consent to participate

Not applicable.

Author details

[1]North Carolina Museum of Natural Sciences, Raleigh, North Carolina 27601, USA. [2]Department of Biostatistics, Yale University, New Haven, Connecticut 06510, USA. [3]Center for Infectious Disease Modeling and Analysis, Yale School of Public Health, Yale University, New Haven, Connecticut 06510, USA. [4]Department of Ecology and Evolutionary Biology, Yale University, New Haven, Connecticut 06525, USA. [5]Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut 06511, USA.

Received: 11 May 2016 Accepted: 24 November 2016
Published online: 01 December 2016

References

1. Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, et al. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science. 2014;346:1320–31.
2. Near TJ, Dornburg A, Eytan RI, Keck BP, Smith WL, Kuhn KL, et al. Phylogeny and tempo of diversification in the superradiation of spiny-rayed fishes. Proc Natl Acad Sci U S A. 2013;110:12738–43.
3. Prum RO, Berv JS, Dornburg A, Field DJ, Townsend JP, Lemmon EM, et al. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. Nature. 2015;526:569–73.
4. Rannala B, Yang Z. Phylogenetic inference using whole genomes. Annu Rev Genomics Hum Genet. 2008;9:217–31.
5. Philippe H, Hervé P, Henner B, Lavrov DV, Littlewood DTJ, Michael M, et al. Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough. PLoS Biol. 2011;9:e1000602.
6. Romiguier J, Ranwez V, Delsuc F, Galtier N, Douzery EJP. Less is more in mammalian phylogenomics: AT-rich genes minimize tree conflicts and unravel the root of placental mammals. Mol Biol Evol. 2013;30:2134–44.
7. Salichos L, Rokas A. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature. 2013;497:327–31.
8. Townsend JP. Profiling phylogenetic informativeness. Syst Biol. 2007;56:222–31.
9. Townsend JP, Su Z, Tekle YI. Phylogenetic signal and noise: predicting the power of a data set to resolve phylogeny. Syst Biol. 2012;61:835–49.
10. Townsend JP, Lopez-Giraldez F. Optimal Selection of Gene and Ingroup Taxon Sampling for Resolving Phylogenetic Relationships. Syst Biol. 2010;59:446–57.
11. Townsend JP, Leuenberger C. Taxon sampling and the optimal rates of evolution for phylogenetic inference. Syst Biol. 2011;60:358–65.
12. Klopfstein S, Kropf C, Quicke DLJ. An evaluation of phylogenetic informativeness profiles and the molecular phylogeny of diplazontinae (Hymenoptera, Ichneumonidae). Syst Biol. 2010;59:226–41.
13. Dornburg A, Alex D, Townsend JP, Matt F, Near TJ. Phylogenetic informativeness reconciles ray-finned fish molecular divergence times. BMC Evol. Biol. [Internet]. 2014;14. Available from: http://dx.doi.org/10.1186/s12862-014-0169-0
14. Gilbert PS, Chang J, Pan C, Sobel EM, Sinsheimer JS, Faircloth BC, et al. Genome-wide ultraconserved elements exhibit higher phylogenetic informativeness than traditional gene markers in percomorph fishes. Mol Phylogenet Evol. 2015;92:140–6.
15. Faircloth BC, Chang J, Alfaro ME. TAPIR enables high throughput analysis of phylogenetic informativeness. arXiv. 2012;1202:1215.
16. López-Giráldez F, Townsend JP. PhyDesign: an online application for profiling phylogenetic informativeness. BMC Evol Biol. 2011;11:152.
17. Su Z, Zhuo S, Zheng W, Francesc L-G, Townsend JP. The impact of incorporating molecular evolutionary model into predictions of phylogenetic signal and noise. Frontiers in Ecology and Evolution [Internet]. 2014;2. Available from: http://dx.doi.org/10.3389/fevo.2014.00011
18. Venditti C, Meade A, Pagel M. Phylogenies reveal new interpretation of speciation and the Red Queen. Nature. 2010;463:349–52.
19. Su Z, Townsend JP. Utility of characters evolving at diverse rates of evolution to resolve quartet trees with unequal branch lengths: analytical predictions of long-branch effects. BMC Evol Biol. 2015;15:86.
20. R Core Team. R: A language and environment for statistical computing [Internet]. 2015. Available from: http://www.R-project.org/.
21. Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics. 2004;20:289–90.
22. Revell LJ. phytools: an R package for phylogenetic comparative biology (and other things). Methods Ecol Evol. 2011;3:217–23.
23. Harmon LJ, Weir JT, Brock CD, Glor RE, Challenger W. GEIGER: investigating evolutionary radiations. Bioinformatics. 2008;24:129–31.
24. McCallum E, Weston S, Parallel R. O'Reilly Media, Inc. 2011.

25. Dornburg A, Alex D, Jon M, Beaulieu JM, Eytan RI, Near TJ. The impact of shifts in marine biodiversity hotspots on patterns of range evolution: Evidence from the Holocentridae (squirrelfishes and soldierfishes). Evolution. 2014;69:146–61.
26. Dornburg A, Friedman M, Near TJ. Phylogenetic analysis of molecular and morphological data highlights uncertainty in the relationships of fossil and living species of Elopomorpha (Actinopterygii: Teleostei). Mol Phylogenet Evol. 2015;89:205–18.
27. Near TJ, Dornburg A, Masayoshi T, Suzuki D, Brandley MC, Friedman M. Boom and Bust: Ancient and recent diversification in bichirs (Polypteridae: Actinopterygii), a relictual lineage of ray-finned fishes. Evolution. 2014;68: 1014–26.
28. Pond SLK, Frost SD, Muse SV. HyPhy: hypothesis testing using phylogenies. Bioinformatics. 2005;21(5):676–9. doi:10.1093/bioinformatics/bti079.
29. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York: Springer; 2016.
30. Jukes TH, Cantor CR. Evolution of Protein Molecules. Mammalian Protein Metabolism. 1969. p. 21–132.
31. Dornburg A, Santini F, Alfaro ME. The influence of model averaging on clade posteriors: an example using the triggerfishes (Family Balistidae). Syst Biol. 2008;57:905–19.
32. Sullivan J, Jack S, Paul J. Model Selection in Phylogenetics. Annu Rev Ecol Evol Syst. 2005;36:445–66.
33. Near TJ, Eytan RI, Dornburg A, Kuhn KL, Moore JA, Davis MP, et al. Resolution of ray-finned fish phylogeny and timing of diversification. Proc Natl Acad Sci U S A. 2012;109:13698–703.
34. Hwang J, Jonathan H, Qi Z, Yang ZL, Zheng W, Townsend JP. Solving the ecological puzzle of mycorrhizal associations using data from annotated collections and environmental samples - an example of saddle fungi. Environ Microbiol Rep. 2015;7:658–67.